International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

A Comprehensive Study of Multimodal Approaches for Medical Visual Question Answering

Tentu Ganga Bhavani

B. Tech Student, Department of CSE, GMR Institute of Technology, Rajam-532127, Andhra Pradesh, India
Email: 21341A05I2@gmrit.edu.in

ABSTRACT

Medical Visual Question Answering is a combination of medical artificial intelligence and popular visual question answering challenges. Medical Visual Question Answering is a multimodal task that uses deep learning to answer clinical questions about medical images. The study aims to answer medical questions based on the visual content of radiology images using deep learning. Deep learning is a method in artificial intelligence (AI) that teaches computers to process data in a way that is inspired by the human brain. Medical Visual Question Answering aims to create an artificial intelligence framework that can efficiently connect medical images and queries, facilitating effective information retrieval. First, this study employed image-to-question (I2Q) consideration to model the relationships between the question and both the visual and linguistic contents. In particular, we add the search query using textual encoded data and extract image features utilizing multimodal. We concatenate the resulting visual and textual representations and feed them into a multi-modal system for generating the answer. Deep learning, medical imaging, and natural language processing are coming together in a promising way with Medical Visual Question Answering, which has the potential to change many facets of healthcare, research, and education.

Keywords: Medical Visual Question Answering; Deep Learning; Multimodal; vision transformer model; encoder transformer.

INTRODUCTION:

Medical Visual Question Answering (Med-VQA) is a field that focuses on answering natural language questions about medical images. Its aim is to provide accurate and convincing answers to clinically relevant questions. VQA is an improved version of Natural Language Question Answering (NLQA), where we give a natural language question to the machine and, based on the knowledge base provided, the machine generates a natural language answer. In this paper, the VQA system discusses medical images and answers questions based on the modality of the image. It can identify image modalities such as X-ray, computed tomography, ultrasound, magnetic resonance imaging, mammography, angiogram, gastro-intestinal imaging, and positron emission tomography. The process of the VQA task can be divided into three parts: extracting image features, extracting question features, and integrating features. For image enhancement, commonly used convolutional neural networks (CNNs) pre-trained on ImageNet include VGGNet, ResNet, and GoogLeNet. Question featurization techniques explored include bag-of-words and LSTM encoders. For doctors, Med-VQA systems can be used to assist diagnosis by providing them with a second medical opinion. The systems can also be used in clinical education to train medical professionals.

RELATED WORKS:

Wang, B et al.,[1] In the biomedical field, pre-trained language models (PLMs) are systematically surveyed in this paper. By standardizing terminology and benchmarks, it seeks to close the gap between various communities. It also addresses constraints and emerging trends to stimulate further research in the field of biomedical PLM. The study evaluates pre-trained language models (PLMs) in the biomedical domain, including PubMedQA, BioASQ, MEDIQA, emrQA, cMedQA, and COVID19-QA, using datasets from MIMIC III, CPRD, BREATHE, and PubMed.

Lubna, A et al.,[2] Using the ImageCLEF 2019 medical VQA dataset, the paper describes the creation of a modality-based medical image visual question answering (VQA) system. The method employed in the paper uses a convolutional neural network (CNN) to determine the answer based on the CNN's output after classifying the input image into its appropriate modality class. First, natural language processing (NLP) methods are used to process the input question. The suggested model's testing accuracy, according to the paper, is 83.8%, which is on par with the state of the art.

Ren, F et al.,[3] In order to tackle the challenging task of medical visual question answering, the paper suggests a model named CGMVQA, which combines classification and answer generation capabilities. A pre-trained ResNet152 mode is used to extract features from images. The authors conducted their investigations using the ImageCLEF 2019 VQA-Med dataset, which consists of 3200 medical images for training, 500 images for validation, and 500 images for testing. In the ImageCLEF 2019 VQA-Med dataset, the CGMVQA model produces cutting-edge results in terms of semantic similarity, word matching, and classification accuracy.

Ambati, R et al.,[4] The study suggests a novel method that combines machine translation and picture captioning techniques for the Visual Question Answering (VQA) task on the medical dataset. The study suggests a sequence-to-sequence model approach that combines machine translation and picture captioning methods for visual question answering (VQA) in the medical field. The study makes use of a dataset that ImageCLEF made available for the ImageCLEF 2018 VQA-med challenge.

Abacha, A. B et al.,[5] The paper details the involvement of the National Library of Medicine (NLM) of the United States in the ImageCLEF 2018 visual question answering task (VQA-Med). Based on the visual content of the images, the task entails responding in natural language to questions regarding medical images. The authors used deep learning networks, specifically multimodal compact bilinear pooling (MCB) and stacked attention network (SAN), for the visual question answering (VQA) task in the medical domain. The study made use of the VQA-Med challenge dataset, which is one of the datasets made available for the visual question answering (VQA) task in the medical domain.

Bar, Y et al.,[6] The paper investigates the efficacy of deep learning techniques for pathology identification in data from chest radiographs. The study uses a deep learning approach based on convolutional neural networks (CNNs), specifically the 5th, 6th, and 7th layers (Decaf 5), as well as SVM with a linear kernel for pathology detection in chest radiograph data. The authors use ImageNet, a sizable non-medical image database with over a million images divided into 1000 categories and a dataset of 93 chest x-ray images, to train their convolutional neural network (CNN).

Hasan, S. A et al.,[7] An overview of ImageCLEF 2019's Medical Visual Question Answering Task (VQA-Med), which focuses on providing medical information based on the visual content of radiology images, is presented in this paper. To complete the task, the authors generated a new dataset at ImageCLEF 2019 consisting of 4,200 radiology images and 15,292 question-answer pairs covering four categories of clinical questions: modality, plane, organ system, and abnormality. With an accuracy of 62.4% and a BLEU score of 64.4%, the top-performing team was attained.

Liu, B et al.,[8] The creation of SLAKE, a sizable, knowledge-enhanced, semantically annotated dataset for Med-VQA system training and testing, is presented in the paper. The SLAKE dataset's creation is presented in the paper. The dataset encompasses a variety of modalities (such as CT, MRI, and X-ray), body parts (such as the head, neck, and chest), and question types. It also contains extensive semantic labels annotated by skilled medical professionals. With a 75:15:15 ratio for each of the eight categories—"head CT," "head MRI," "neck CT," "chest X-Ray," "chest CT," "abdomen CT," "abdomen MRI," and "pelvic cavity CT"—the SLAKE dataset was divided into training, validation, and test sets at the image level.

Zhu, C et al.,[9] For visual question answering tasks, the paper suggests an organized attention mechanism that outperforms baseline models on the CLEVR and VQA datasets. Three datasets—the SHAPES dataset, the CLEVR dataset, and the VQA real-image dataset—were used for evaluation, according to the paper. The model performed 9.5% better on the CLEVR dataset than the best baseline model and 1.25 percent better on the VQA dataset than the best published model.

Kafle, K et al.,[10] The current state of Visual Question Answering (VQA) is critically examined in this paper with regard to problem formulation, existing datasets, evaluation metrics, and algorithms. In addition to discussing multi-modal low-rank bilinear pooling (MLB) scheme, the paper addresses the use of Multimodal Compact Bilinear (MCB) pooling as a method for combining image and text features in Visual Question Answering (VQA). The VQA Dataset contains images from the COCO dataset (COCO-VQA), which includes DAQUAR, FM-IQA, Visual7W, Visual Genome, and The VQA Dataset, in addition to synthetic cartoon imagery (SYNTH-VQA)

Wu, Q et al.,[11] This paper offers an extensive overview of datasets and methodologies in the Visual Question Answering (VQA) field. By contrasting contemporary solutions to the VQA problem, it assesses the state of the art. Four categories of VQA methods are presented in the paper: knowledge base-enhanced approaches, compositional models, attention mechanisms, and joint embedding approaches. Images from the NYU-Depth v2 dataset were used to create the DAQUAR dataset. There are 12,468 question-answer pairs gathered from both artificial and human annotations on 795 training and 654 test images. Images from the Microsoft Common Objects in Context (COCO) dataset, which has over 2 million labeled instances, 91 common object categories, and 328,000 images, are among the datasets.

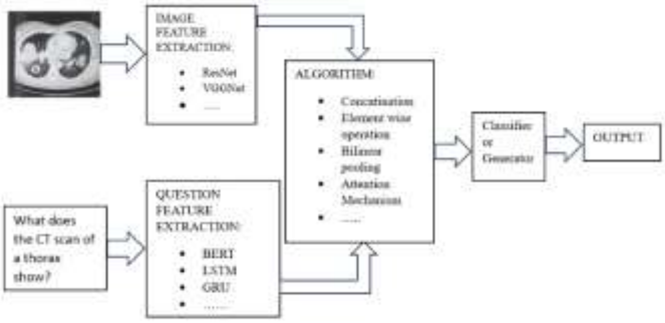
Wang, J et al.,[12] The performance of Text-VQA methods is hampered by the lack of human-labeled question-answer (QA) pairs, a problem that is addressed in this paper. For the Text-VQA task, the paper uses the TextVQA dataset and the ST-VQA dataset. Like BERT BASE, the multi-modality fusion module in TAG is a four-layer transformer with twelve attention heads. The TAG method was evaluated both qualitatively and quantitatively on the TextVQA and ST-VQA datasets.

Zhang, K et al.,[13] The study presents BiomedGPT, a general-purpose Biomedical Generative Pre-trained Transformer model that can accept multi-modal inputs and carry out a range of downstream operations. The BART architecture, which is used as a sequence-to-sequence model with a BERT-style encoder over corrupted text and a GPT-style left-to-right autoregressive decoder, serves as the foundation for BiomedGPT's design. The paper reports results on the MedMNIST v2 dataset, a collection of benchmark datasets spanning multiple biomedical domains, for the image classification task.

Melvin, Y. J et al.,[14] The purpose of this paper is to apply data mining techniques to address the problems associated with Visual Question Answering (VQA) for medical images, particularly when dealing with clinic trial text in English and multilingual contexts. Inception Resnet V2 model, Bidirectional LSTM (Bi-LSTM), and Random Forest algorithm are the methods utilized in this paper. This paper makes use of a labeled medical image dataset. A training set and a test set are separated into the dataset, with a 70:30 supervised learning ratio.

Sampat, S. K et al.,[15] In this paper, a novel question-answering task is presented, which requires the participant to mentally simulate the potential outcomes of executing particular actions in a given scenario. In order to propose baseline solvers for this task, the authors modify existing VQA methods and create a vision-language question answering task based on the CLEVR dataset. The CLEVRHYP dataset is used in the paper; it is a synthetic dataset designed for the task of visual question answering using hypothetical actions over pictures.

METHODOLOGY:



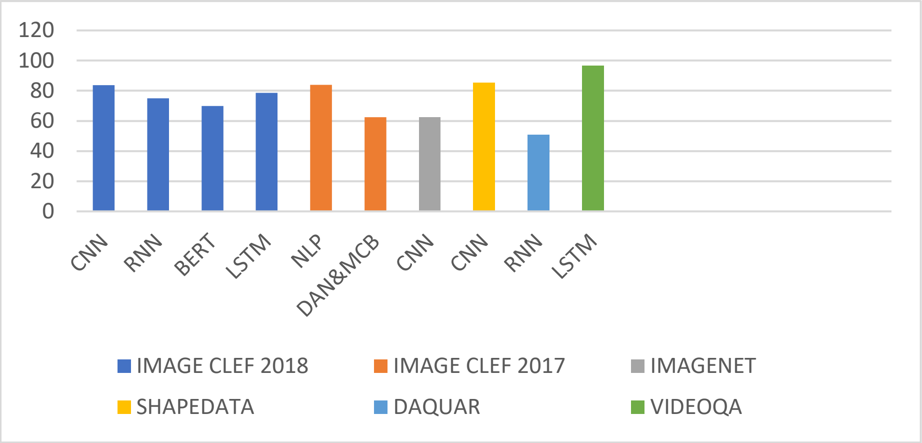
A question encoder, an answering component, a feature fusing algorithm, and an image encoder make up the framework's four main parts. Well-developed convolutional neural network (CNN) backbones like VGG Net and ResNet can be used as the image feature extractor, and popular language encoding models like LSTM and Transformer can be used as the question encoder. During the VQA task training, the feature encoding models can be either frozen or end-to-end fine-tuned. Typically, the models are started with pre-trained weights. Usually, a recurrent neural network language generator or a neural network classifier serves as the answering component. Element-wise multiplication is used in the LSTM to fuse the question and image features. To improve system performance even more, researchers created novel fusing algorithms and included the well-liked attention mechanism.

The image features are extracted using ResNet, VGGNet, etc. RestNet stands for residual network, used in image feature extraction due to its ability to train very deep neural networks efficiently. In ResNet, by removing the fully connected layers and utilizing the convolutional base, we can extract generic features of the image. VGGNet stands for Visual Geometry Group Network, used in image feature extraction. The question features are extracted using LSTM, GRU, Transformer, etc. The output after extracting the image feature and question feature is combined using some algorithms like concatenation, element-wise multiplication, bilinear pooling, attention mechanism, etc. The choice of concatenation, element-wise multiplication, bilinear pooling, or attention mechanism depends on the characteristics of the data, model complexity, and task requirements. Then we apply a classifier or generator to generate the output.

RESULT:

S.NO	AUTHOR	DATASET	METHOD	ACCURACY	PRECISION	RECALL	F1SCORE	BLEU	WBSS
1	Lubna, A., Kalady, S., & Lijiya,	ImageCLEF 2019 VQA-Med	NLP and CNN	83.8%	80.45%	81.28%	79.66%	-	-
2	Ren, F., & Zhou	ImageCLEF VQA-Med	Data Agumentat -ion and Tokenization ,Pre-trained ResNet152	64.0%	64.3%	64.6%	62.6%	-	-
3	Ambati, R.,& Dudyala, C. R.	ImageCLEF VQA-Med	Pre-trained VGG16 network, Question encoder, answer	-	-	-	-	0.1326	0.173
4	Abacha, A. et al.,	ImageCLEF 2018 VQA-Med	Deep learning	99.19%	-	-	-	0.1006	0.1546
5	Hasan, S. A., Ling,	ImageCLEF 2018 VQA-Med	CNN	62.4%	-	-	-	0.1204	0.1402

6	Ben Abacha, A., Hasan	ImageCLEF 2019 VQA-Med	NLP and Computer Vision techniques	51.56%	-	-	-	0.5349	-



CONCLUSION:

This presents a survey of the dataset and approaches to medical visual question answering. The field of Visual Question Answering (VQA) has advanced significantly, but its use in the medical field is still in its beginning. The four main challenges that the medical VQA system faces are as follows: First, how can the system respond to a range of deeper question categories? The second question is how to incorporate medical features into the task; the third is how to validate a response to increase its conclusiveness; the fourth is how to prevent bias in the system towards any modality; and the fifth and final question is how to optimize the VQA for medicine during the process. The objective is to stimulate and motivate additional research, resulting in developments that support the expansion and ability of medical VQA within the framework of artificial intelligence and healthcare. The majority of the models produce results based on the classification task; evaluating them on the medical dataset presents a challenge. MedVQA systems can be used as educational tools for medical students and healthcare professionals, offering interactive learning experiences by answering questions related to medical images. It can assist researchers in mining large datasets of medical images, helping to extract valuable insights and patterns for scientific studies and clinical research.

References

1. Wang, B., Xie, Q., Pei, J., Chen, Z., Tiwari, P., Li, Z., & Fu, J. (2021). Pre-trained language models in biomedical domain: A systematic survey. ACM Computing Surveys.
2. Zhang, K., Yu, J., Yan, Z., Liu, Y., Adhikarla, E., Fu, S., ... & Sun, L. (2023). BiomedGPT: A Unified and Generalist Biomedical Generative Pre-trained Transformer for Vision, Language, and Multimodal Tasks.
3. Lubna, A., Kalady, S., & Lijiya, A. (2019). Mobvqa: A modality based medical image visual question answering system, TENCON 2019-2019.
4. Ren, F., & Zhou, Y. (2020). Cgmvqa: A new classification and generative model for medical visual question answering. IEEE Access, 8, 50626-50636.
5. Ambati, R., & Dudyala, C. R. (2018, December). A sequence-to-sequence model approach for imageclef 2018 medical domain visual question answering. In 2018 15th IEEE India Council International Conference (INDICON) (pp. 1-6). IEEE.
6. Abacha, A. B., Gayen, S., Lau, J. J., Rajaraman, S., & Demner-Fushman, D. (2018, September). NLM at ImageCLEF 2018 Visual Question Answering in the Medical Domain.
7. Bar, Y., Diamant, I., Wolf, L., & Greenspan, H. (2015, March). Deep learning with non-medical training used for chest pathology identification. In Medical Imaging 2015: Computer-Aided Diagnosis (Vol. 9414, pp. 215-221). SPIE.
8. Hasan, S. A., Ling, Y., Farri, O., Liu, J., Müller, H., & Lungren, M. (2018). Overview of imageclef 2018 medical domain visual question answering task.
9. Liu, B., Zhan, L. M., Xu, L., Ma, L., Yang, Y., & Wu, X. M. (2021, April). Slake: A semantically-labeled knowledge-enhanced dataset for medical visual question answering. In 2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI) (pp. 1650-1654). IEEE.
10. Zhu, C., Zhao, Y., Huang, S., Tu, K., & Ma, Y. (2017). Structured attentions for visual question answering. In Proceedings of the IEEE International Conference on Computer Vision (pp. 1291-1300).

11. Kafle, K., & Kanan, C. (2017). Visual question answering: Datasets, algorithms, and future challenges. Computer Vision and Image Understanding, 163, 3-20.

12. Wu, Q., Teney, D., Wang, P., Shen, C., Dick, A., & Van Den Hengel, A. (2017). Visual question answering: A survey of methods and datasets. Computer Vision and Image Understanding, 163, 21-40

13. Wang, J., Gao, M., Hu, Y., Selvaraju, R. R., Ramaiah, C., Xu, R., ... & Davis, L. S. (2022). Tag: Boosting text-vqa via text-aware visual question-answer generation. arXiv preprint arXiv:2208.01813.

14. Melvin, Y. J., Gawade, S., & Palivela, H. (2021, March). Visual Question Answering using Data Mining Techniques for Skeletal Scintigraphy in medical domain-VQADMSS. In 2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS) (pp. 859-863). IEEE.

15. Sampat, S. K., Kumar, A., Yang, Y., & Baral, C. (2021). CLEVR_HYP: A challenge dataset and baselines for visual question answering with hypothetical actions over images. arXiv preprint arXiv:2104.05981.